Thursday Exam 1

### Background

You are working as a marketing analyst for a professional basketball team seeking to improve **ticket sales** through more effective promotional strategies. Over the course of a season, the team has run 100 marketing campaigns and recorded a variety of data points for each. These campaigns span digital outreach, physical advertising, and fan engagement tactics. Your job is to analyze the data and develop a regression model that helps predict total sales (sales) based on these different forms of marketing activity.

The variables collected for each campaign include: the number of email\_clicks generated from newsletters, the number of social\_media\_mentions observed across platforms, the amount of ad\_spend used on digital platforms, the number of merch\_giveaways distributed during events, the estimated billboard\_exposure in terms of views, and the number of tv\_spots aired during peak hours. Using this dataset, you will build and compare several linear regression models to assess which combinations of marketing efforts most effectively predict ticket sales. This data was created based on simulated data.

### Data Generation

# A tibble: 7 × 3  
 variable min max  
 <chr> <dbl> <dbl>  
1 ad\_spend 56.8 800.  
2 billboard\_exposure 139. 371.  
3 email\_clicks 188. 418.  
4 merch\_giveaways 7.31 577.  
5 social\_media\_mentions 84.8 439.  
6 ticket\_sales 309. 680.  
7 tv\_spots 20.3 190.

### Model Building

## Model 1  
model1 <- lm(ticket\_sales ~ email\_clicks + social\_media\_mentions, data = marketing\_df)  
  
## Model 2  
model2 <- lm(ticket\_sales ~ email\_clicks + social\_media\_mentions + ad\_spend, data = marketing\_df)  
  
## Model 3  
model3 <- lm(ticket\_sales ~ merch\_giveaways + billboard\_exposure + tv\_spots, data = marketing\_df)

model\_1\_summary <- summarize\_reg\_model(model1,"Model 1")  
model\_2\_summary <- summarize\_reg\_model(model2,"Model 2")  
model\_3\_summary <- summarize\_reg\_model(model3,"Model 3")  
  
bind\_rows(  
 model\_1\_summary,  
 model\_2\_summary,  
 model\_3\_summary  
)

type RSS RSE R2 Adj\_R2 AIC BIC  
1 Model 1 822565.1 28.72 0.73 0.73 9558.30 9577.94  
2 Model 2 795082.3 28.25 0.74 0.74 9526.32 9550.86  
3 Model 3 1396253.1 37.44 0.55 0.55 10089.42 10113.96

Note:

Model 1 - ticket\_sales ~ email\_clicks + social\_media\_mentions

Model 2 - ticket\_sales ~ email\_clicks + social\_media\_mentions + ad\_spend

Model 3 - ticket\_sales ~ merch\_giveaways + billboard\_exposure + tv\_spots

**An employee at the company suggested that model 2 should be used for predicting ticket sales.**

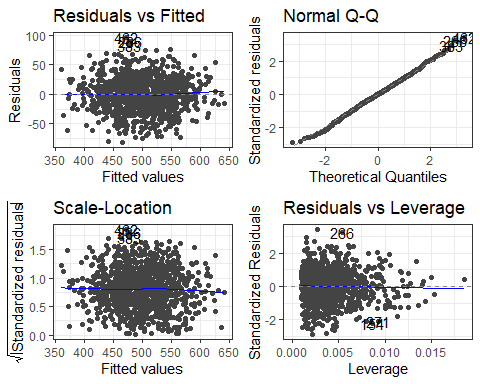
summary(model2)

Call:  
lm(formula = ticket\_sales ~ email\_clicks + social\_media\_mentions +   
 ad\_spend, data = marketing\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-80.994 -17.948 0.094 18.746 95.850   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.191e+02 7.135e+00 16.689 < 2e-16 \*\*\*  
email\_clicks 1.172e+00 2.436e-02 48.090 < 2e-16 \*\*\*  
social\_media\_mentions 3.553e-02 1.350e-02 2.632 0.00863 \*\*   
ad\_spend 5.249e-02 8.946e-03 5.868 6.02e-09 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 28.25 on 996 degrees of freedom  
Multiple R-squared: 0.7434, Adjusted R-squared: 0.7426   
F-statistic: 961.9 on 3 and 996 DF, p-value: < 2.2e-16

anova(model1,model2)

Analysis of Variance Table  
  
Model 1: ticket\_sales ~ email\_clicks + social\_media\_mentions  
Model 2: ticket\_sales ~ email\_clicks + social\_media\_mentions + ad\_spend  
 Res.Df RSS Df Sum of Sq F Pr(>F)   
1 997 822565   
2 996 795082 1 27483 34.428 6.017e-09 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Perform residual analysis plot  
autoplot(model2) + ## from ggfortify  
 theme\_bw()



resettest(model2) # From lmtest package

RESET test  
  
data: model2  
RESET = 0.66602, df1 = 2, df2 = 994, p-value = 0.514

dwtest(model2) # From lmtest package

Durbin-Watson test  
  
data: model2  
DW = 1.9835, p-value = 0.3967  
alternative hypothesis: true autocorrelation is greater than 0

shapiro.test(resid(model2)) # Base R

Shapiro-Wilk normality test  
  
data: resid(model2)  
W = 0.99845, p-value = 0.5245

bptest(model2) # From lmtest package

studentized Breusch-Pagan test  
  
data: model2  
BP = 4.5622, df = 3, p-value = 0.2068

**Model 2: VIF**

vif(model2) ## From car package

email\_clicks social\_media\_mentions ad\_spend   
 1.111806 1.137386 1.221817

**Model 2: Partial Plots**

avPlots(model2) ## From car package

